DOCUMENT RESUME

ED 156 722 TM 007 378

AUTHOR Edwards, Keith J.
TITLE Fair Employment and Performance Appraisal: Legal Requirements and Practical Guidelines.
PUB DATE 5 Sep 76
NOTE 34p.; Paper presented at the Annual Meeting of the American Psychological Association (84th, Washington, D.C., September 5, 1976) ; Not available in hard copy due to poor reproducibility

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *Civil Rights Legislation; Court Litigation; Criterion Referenced Tests; Employers; *Employment Problems; *Equal Opportunities (Jobs); Evaluation Criteria; Guidelines; Job Analysis; Job Applicants; *Legal Problems; Legal Responsibility; Minority Groups; *Occupational Tests; Personnel Evaluation; Personnel Selection; Predictive Validity; Testing Problems; *Test Validity
IDENTIFIERS *Civil Rights Act 1964 Title VII; Judicial Validity

ABSTRACT

The use of tests in personnel decisions has become an increasing legal liability for employers. The major questions raised by the courts concerning this use of tests are described. Current federal guidelines for performance appraisal systems, as established by the Equal Employment Opportunity Commission, are explained and traced to Title VII of the 1964 Civil Rights Act. The legal implications of prima facie discrimination and the assessment of adverse impact upon minorities is explained. The process of judicial validation of performance appraisal systems is discussed, including specific case examples and flow charts. In deciding whether the criteria used in personnel decisions are valid and nondiscriminatory, the courts have utilized predictive, concurrent and content validity evaluations. The legal preferences and problems associated with each type of validity are described. The role of selection ratios, adverse impact, and business necessity in judicial validity decisions is discussed. The courts are concerned with the statistical correlation between test results and the criterion measures of job performance, and it is felt that the conflicting definitions of test validity and fairness provided by industrial psychologists have caused problems in the courts. The social controversy surrounding civil rights and employment testing is also discussed briefly. (Author/JAC)

FAIR EMPLOYMENT AND PERFORMANCE APPRAISAL: LEGAL REQUIREMENTS AND PRACTICAL GUIDELINES

KEITH J. EDWARDS

Rosemead Graduate School of Psychology
Rosemead, CA 91770

Title VII of the 1964 Civil Rights Act required that no employer discriminate against any individual on the basis of race, color, creed, sex, or national origin. Since the inactment of this legislation some 12 years ago, personnel practices of both public and private employers have been subject to severe scrutiny. The results of well over 100 legal cases has clearly demonstrated that personnel practices long accepted as routine and essential can become significant legal liabilities. The issues involved have been emotionally charged and the stakes have been high. For example, over a period of more then a year during 1975-76 Chicago was without 75 million to 96 million dollars in revenue sharing funds. A federal judge had ordered the funds impounded and declared the city's police officer's exam, sergeant's exam, and sergeant's performance ratings discriminatory against blacks and chicanos in violation of Title VII of the Civil Rights Act. The city was enjoined from further use of the tests or ratings and ordered to hire and promote minorities and women in accordance with court imposed quotas until such time as nondiscriminatory testing could be established (USA v. City of Chicago, et al January 5, 1976, Memorandum Decision).

A paper presented on the symposium "Performance Appraisal and Feedback: Flies in the Ointment", Division 14, Annual Meeting of the American Psychological Association, Washington, D. C., September 5, 1976. An earlier version of paper was presented at the Conference on Performance Appraisal, Center reative Leadership, Greensboro, N.C., January, 1976.

On January 18, 1973 the U.S. District Court for the Eastern District of Pennsylvania approved a consent decree between AT & T, the Equal Employment Opportunity Commission, and the Department of Labor. The company agreed to compensate women and minority employees with payments which were estimated to run between 12 and 13 million dollars. The payments were intended as retroactive compensation to those who in the past may have been victims of discrimination in promotion, transfers, and salary administration (Miner, 1974). In addition AT & T agreed to implement personnel practices which would achieve a balance between the proportions of women and minorities in its various occupations and proportions in the relevant labor force. The latter type of agreement is referred to in various government memoranda as voluntary goals and timetables as distinguished from the Court imposed quotes exemplified in the Chicago decision. A further note about the AT & T consent decree is that the company agreed that results of future testing of minority applicants could not be used as a justification for failure to achieve the goal of proportional representation.

In light of such outcomes of litigation, one can understand why employers are discontinuing the use of testing and performance appraisals for hiring and promotion. They have assessed the uncertainties of the current situation and decided that whatever gain in organizational efficiency these personnel practices provide is not worth the legal risks involved. The inevitable result has been movement toward random hiring and promotion. However, the long term effects of loss of efficiency when valid employer assessment procedures are discontinued is also a high price to pay. Thus, the employer

who is faced with weighing the legal risks of keeping a performance appraisal system versus the long term inefficiency risks of dropping it needs to be able to properly assess the strengths and weaknesses in their system. The purpose of the present paper is to outline the major legal questions courts have raised concerning the use of tests and other criteria in personnel decision-making. The paper will be primarily concerned with the legal questions raised concerning appraisal systems. Also, the concept of judicial validation will be defined. By judicial validation I mean the decision process which has developed in the courts to evaluate whether criteria used to make personnel decisions are valid or invalid, nondiscriminatory or discriminatory, legal or illegal.

One might question the wisdom and utility of conceptualizing yet another type of validity. What with content, construct, concurrent, predictive, face, and synthetic validities already in the psychologist's lexon, who needs another? I would argue that judicial validity has much to recommend itself not the least of which is that it is the one that counts in the real world.

Origins of Current Federal Guidelines

In understanding the legal requirements imposed on performance appraisal systems it is helpful to consider current requirements in historical perspective. Title VII of the 1964 Civil Rights Act addresses, in very general terms, the notion that it is unlawful for employers to discriminate on the basis of race, color, religion, sex, and national origin. In the course of congressional debate the Tower Amendment was added to the bill in an attempt

to clarify the intent of the Act. Briefly the Tower Amendment said the Act did not preclude an employer's use of and acting on any professionally developed test, provided such test is not designed, intended, or used to discriminate because of race, color, religion, sex or national origin. Originally Senator Tower formulated the Amendment to forestall the intrusion of the Federal Government into management's right to prescribe employment qualification (Ash, 1966). In retrospect, the Amendment has not had such an effect and was in fact the main stimulus for current federal guidelines which detail the characteristics of acceptabel personnel decision systems.

In 1966, the Equal Employment Opportunity Commission, as the administrative agency charged with implementing the provision of the Civil Rights Act, issued its first set of guidelines, titled "Guidelines on Employment Testing Procedures." It is clear from the first paragraph in the 1966 EEOC Guidelines that they were formulated to interpret the language in the Tower Amendment. Two major characteristics of the 1966 guidelines were the adoption of the APA Standards for Educational and Psychological Tests and Manuals (1966) and the requirement of separate criterion-related validity studies for minority and majority groups. This latter requirement involves checking for differential validity which will be explained in more detail later on in the paper.

In 1970, the EEOC published a revised version of the guidelines in the Federal Register. The title on the revision is "Guidelines on Employee Selection Procedures." The word testing in the title of the 1966 guidelines had been changed to selection which suggests the expanded scope of the 1970 guidelines. Of major concern for the present discussion is the comprehensive definition of the word "test" in the 1970 guidelines:

> For the purpose of the guidelines in this part, the term 'test' is defined as any paper and pencil or performance measure used as a basis for any employment decision... The term 'test' includes all formal scored, quantified or standardized techniques of assessing job suitability...(1607.2).

In a recent issue of The Conference Board Records (Lazer, 1976) offered a tentative conclusions that this definition covers performance appraisals. It is clear from the scope of this definition and subsequent interpretations by the courts that use of performance appraisal systems for employment decisions comes under these guidelines. It is the interpretation by the courts of the 1970 EEOC guidelines which provide the current legal definition of nondiscriminatory personnel practices.

EEOC Guidelines

In attempting to understand the EEOC guidelines one must remember they were originally formulated to cover the use of paper and pencil ability tests or "professionally developed tests" in the words of the Tower Amendment. In the process, EEOC has adopted as a minimum standard, the standard for test validity set forth by the American Psychological Association. There are many who are of the opinion that applying standards developed for commercially produced written tests to systems of assessment like performance appraisals makes it extremely difficult or impossible to defend performance appraisals. The lack of success employers have had in defending performance appraisals in the courts seems to substantiate this conclusion. I would hasten to add, however, that the performance appraisal systems challenged in court cases to date have not been sterling examples of sound

personnel practice. (cf Holley and Field, 1975 for a review of several cases involving performance appraisals). This fact has made defense of good appraisal systems even more difficult. The requirements for demonstrating that an appraisal system is nondiscriminatory have become more complex and more stringent. Further, the courts have developed a deep scepticism about any assessment technique involving supervisory judgments. In fact, the District Court in the Chicago Police Department case in January of this year concluded without qualification "that supervisory ratings are not a fair measurement of an employee's suitability for promotion " (US v. City of Chicago, 8 EDP 9785, 1974). The court further interprets the testimony of defendents' and plaintiffs' expert witnesses, both well known industrial psychologists, as being in agreement with this conclusion. With such an unqualified negation of the usefulness of supervisor ratings, it is little wonder that defense of any performance appraisal system, no matter how thoroughly developed, is an uphill battle.

Most of the court decisions to date have not had as their major focus, the question of the validity of performance appraisals. However, supervisor ratings have been a favorite criterion in predictive and concurrent validation studies. From the courts critiques of such studies one can garner a great deal of information on how they view supervisor's ratings. For example, the Supreme Court in Albemarle v. Mood found the validation studies conducted by the employer materially defective in part because "Albemarle's supervisors were asked to rank employees by a 'standard' that was extremely vague and fatally open to divergent interpretation..." Lower courts have manifested a more general suspicion of supervisor ratings as exemplified by the above quote from the Chicago Police Department case.

What then are the legal requirements for a performance appraisal system to be nondiscriminatory? The definition of discrimination given in the EEOC guidelines was endorsed by the Supreme Court in the 1971 Griggs vs Duke Power case as "the administrative interpretation of the (Civil Rights) Act by the enforcing agency" and consequently entitled to "great deference." The following is the complete text of the EEOC definition of discriminatory use of "tests":

> The use of any test which adversely affects hiring, promotion, transfer or any other employment or membership opportunity of classes protected by title VII constitutes discrimination unless; (a) the test has been validated and evidences a high degree of utility as hereinafter described, and (b) the person giving or acting upon the results of the particular test can demonstrate that alternative suitable hiring, transfer or promotion procedures are unavailable for his use. (1607.3).

In the decision of the Supreme Court in Albermarle vs Moody in June 1975, the court reaffirmed its endorsement of part (a) of the EEOC definition but modified part (b). It is now legally the burden of the complaining party to make a showing that other procedures for hiring, transfer or promotion are available.

The first part of the definition involves what the courts have come to define as a prima facie case of discrimination. Specifically, performance appraisals are prima facie discriminatory if their use in personnel decision making results in hiring, promotion, transfer, or layoffs in a racial pattern significantly different from the pool of applicants. The burden of proving a prima facie case of discrimination lies legally with the complaining party. If plaintiffs can demonstrate that decisions based upon performance appraisals have an adverse impact on minorities, the burden of proof for establishing the validity of the performance appraisals is shifted to the employer. Thus,

the process of proving a charge of discrimination involves two steps: first, the plaintiffs must establish a prima facie case of discrimination involving adverse impact on minorities; second, the employer must fail to demonstrate a relationship between performance appraisal scores and performance on the job. While the first step in the process is the legal burden of complaining parties, it behoves employers to make careful assessment of any adverse impact on minorities decisions based upon performance appraisals are having.

Assessing Adverse Impact

The means by which one assesses the adverse impact of performance appraisals depends on the nature of the personnel decision it supports. If the decision is dichotomous, such as promote or not promote, retain or lay-off, then a direct statistical comparison of proportions of minority and majority applicants assigned to the same status should be made. If the performance appraisal results in assigning employees to categories such as more than acceptable, acceptable, questionable, and not acceptable, then statistical comparisons of the frequencies of minorities and nonminorities in each category should be made. If a numerical score is assigned to individuals such as in the use of summated ratings or behaviorally anchored scales, then the averages for minority and nonminority groups should be statistically compared. In each of these comparisons, if the differences observed are likely to occur less than once in twenty times by chance alone, the courts are certain to consider this clear evidence of adverse impact.

While statistically significant differences between performance of minority and nonminority groups is sufficient to establish a prima facie case, it is not always necessary. Lower courts, in interpreting the

Supreme Court's position on what constitutes a prima facie case, have approved other ways to establish such a showing. Extreme under-representation of minorities in various eschelons of a promotional structure may establish a prima facie case without reference to applicant pools. Disparities between proportions of minorities employed by a company compared to the general population or a similarly situated work force may also establish a prima facie case. Finally, apparent discrepancies between the performance ratings of minority and nonminorities not statistically significant may still be interpreted by the courts as adverse impact. The Chicago police case cited earlier involved efficiency ratings with one point difference in means of 85.2 for whites and 84.3 for blacks. The court considered the difference to be significant because 92% of all patrolmen scored between 80 and 95 on the measure.

Clearly, the first major question an employer should address concerns the relative effects of performance appraisals on minorities and non-minorities. If scores produced by such appraisals or decisions based upon these scores do not adversely affect minorities, the appraisals are by definition nondiscriminatory. All other things being equal, the performance appraisal system which minimizes differences between minorities and non-minorities has the least legal liability under Title VII as interpreted by the EEOC guidelines. If performance appraisals do not have adverse impact, the employer has no legal burden of proving the appraisal scores are related to job performance. If there is adverse impact, the performance appraisals are prima facie discriminatory and the employer must present empirical evidence to the courts proving the appraisals are valid.

Before turning to the requirements for proving validity, I would note that the employer is responsible to keep accurate records on the results of employment decisions for each protected group under Title VII. At one time it was not considered proper or even legal to keep minority identification in the personnel records. Under Title VII requirements such information is essential.

Establishing Validity

The 1970 EEOC guidelines specify the following requirement for establishing the validity of a "test":

> Evidence of a test's validity should consist of empirical data demonstrating that the test is predictive of or significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which the candidates are being evaluated (1607.4(c))

As noted earlier, the 1966 EEOC guidelines were patterned after the APA standards applicable to written tests used mainly in hiring new employees. The 1970 guidelines were extended to apply to virtually every criteria which could be used as a basis in making personnel decisions. The extension came solely in the broader definition of a test cited earlier. The requirements for evidence of validity, which constitute the bulk of the 1970 version, endorse the concept of criterion-related validity developed for acievement or ability tests.

Criterion-related validity involves the process of correlating individual test scores to independent measures of actual job performance. There are two types of criterion-related validity, predictive and concurrent. Predictive validity is longitudinal in design. Applicants are tested before being hired

and then followed up, after a period of time on the job, with measures of job performance. Concurrent validity is cross-sectional in design. Job incumbants are given the test and the job performance measures at about the same time. In both cases, correlations between test scores and performance scores constitute the validity evidence.

Courts have shown a distinct preference for predictive over concurrent validity. This is primarily because a position, admission to which is contingent upon a test which has adverse impact, is not likely to have many minorities as job incumbants. Since incumbants constitute the sample of a concurrent study, no stable estimate of the relationship between test scores and job performance measures for minorities can be made. The courts have been consistent in rejecting evidence of test validity based upon nonminority samples as evidence that the test is valid for minorities. The guidelines require that "data must be generated and results separately reported for minority and nonminority groups " (1607.5 (b) (5)).

The requirement that employers must check for possible differential-validity is thoroughly imbedded in the case law of Title VII litigation. Protestations by professional psychologists that differential validity doesn't exist are not likely to remove it in the near future. One of the main controversies surrounding the currently proposed uniform guidelines of the Equal Employment Opportunity Coordinating Council (EEOCC) is whether or not the requirement to check for differential validity should be retained.

The guidelines, while definitely endorsing criterion related validity, are ambivalent concerning the alternative approach to validity called content validity. The best understanding of content validity can be obtained from the definition given it by the courts:

> For a test to be content valid, the knowledge, skills and aptitudes required for successufl examination performance must be the knowledge, skills and aptitudes required for successful job performance (US vs City of Chicago, 8 EPD, 9785)

The definition given in the APA 1974 standard states:

> To demonstrate content validity of a set of test scores, one must show that the behaviors demonstrated in testing constitute a representative sample of behaviors to be exhibited in a desired performance domain (p. 28).

The ambivalence of the EEOC guidelines on content validity is reflected in the following two statements from section 1607.5 (a):

> "Evidence of content validity... may also be appropriate where criterion-related validity is not feasible."

> "Evidence of content-validity alone may be acceptable for well developed tests that consist of suitable samples of the essential knowledge, skills or behavior comparing the job in question."

The first statement says content validity is definitely second best. The second statement provides a qualified endorsement of content validity. The courts have extrapolated EEOC ambivalence to an implicit but clear hierarchy of validity approaches:

Predictive Validity - most preferred

Concurrent Validity - second best

Content Validity - least preferred

The courts have identified content validity as "a form of 'rational validity', rather than empirical' validity. The (content validity) analysis depends appreciably on opinions of psychologists " (US vs City of Chicago, 8 EPD, 9785).

It is my view that this distinction between rational and empirical validity is illusory, more semantic than real. In fact, as one examines the text of court opinions in Title VII cases, it is clear that a comprehensive process of "rational validity" has been developed by the judicial system. This process is what I have called "judicial validation." A limitation on the usefulness of the concept of judicial validity is that, to date, courts have had a great deal more to say about what is not acceptable than what is acceptable employment practice. Perhaps it would better be labeled "judicial invalidity." Whatever term one uses, I mean to refer to the process by which the courts have systematically examined an employment system, critiqued empirical evidence, and weighed the evidence of expert witnesses in arriving at a decision that a test is invalid or valid. Of specific interest for the present discussion is the process used by the courts to evaluate performance ratings.

I have formulated a decision flow chart as a means of summarizing my conception of judicial validity. I would like to outline by means of the flow chart, the factors which have been emphasized by courts in critically examining performance appraisals. The critical element throughout the chart is V, the judicial validity coefficient. It is a qualitative index which is used to illustrate the relative contribution of each factor to the overall decisions. The weightings involved are based upon my intutitive estimate of relative contribution and are presented as a didactic tool. No precision is implied in the numbers used. However, the relative sizes of the various terms are meant to reflect my perceptions of the relative merits each step has in litigation.

The Judicial Validation Process

What has been covered up to now has been the guidelines formulated as the Supreme Court said in Griggs to be the administrative interpretation of the Civil Rights Act. However, ultimate interpretation of legislation is a judicial responsibility. Judicial interpretations are the ones that count in the real world. What follows is my conception of the court's interpretation of Title VII to date. I would emphasize the qualification to date. Judicial interpretations are dynamic and thus a description at any point in time can quickly become out dated.

Further, I would caution that I have attempted to describe an historical process not a normative one. In many instances my perception of what is and what ought to be are quite divergent, I do not in this context discuss the latter.

Throughout the flow chart which I have used to define judicial validation (JV), the variable (V) is used as a crude quantification of the outcome. Its purpose is primarily heuristic. Also, all decision points have been arbitrarily reduced to a small number of discrete alternatives. Undoubtably, some of the parameters involved are continuous (eg. adverse impact statistics) but such representation would involve unnecessary complexity.

The first and crucial point of JV is the evaluation of adverse impact of the personnel decision the basis for which is some appraisal system. It should be emphasized that decisions have adverse impact, tests do not. Criteria used by the decision maker become the subject of JV only to the

extent that the decisions adversely affect a protected group. Thus it is possible to employ "tests" to make personnel decisions and avoid their legal liability by monitoring their impact. Giving up testing or stopping appraisals is not the only way to avoid litigation. Crudely put if your numbers come out right, you have nothing to worry about. If the decisions have no adverse impact, by definition the criteria used are nondiscriminatory.

If there is adverse impact, the outcome of JV will be a function of its degree. This contingency is not discussed in any of the so called guidelines but it is a legal reality. The Supreme Court in Albemarle v Moody reflected the contingency when it ruled that "...there simply was no way to determine whether the criteria actually considered were sufficiently related to the company's legitimate interest in job specific ability to justify a testing system with a racially discriminatory impact" (p. 305)(emphasis added). Further, it is my personal opinion that the recent ruling of the Court in Washington v Davis was due in part to the mild adverse impact of Test 21.

In making the outcome of JV contingent in part upon the degree of adverse impact, the courts have included social utilities in the validation process. Recent writers such as Peterson and Novick (1976) have proposed the inclusion of such utilities in psychometric models, but we are far from a consensus.

The second decision point is to decide if the adverse impact of the appraisal system is due to seniority. For the first several years in a new position, appraisal scores are likely to be correlated with seniority. It may be that minorities have less seniority and thus lower scores due to

past discrimination. The courts have held that such a situation is the present effect of a past practice of discrimination and have ordered the appraisals dropped or altered to reduce the adverse impact (cf. Harper vs Mayor and City Council of Baltimore).

In the next step of JV we find one of the pivotal issues on which discrimination cases turn--the job analysis. Is the appraisal system based upon a comprehensive job analysis? Did the employer attempt to systematically identify the essential knowledge, skills, and behaviors composing the position being appraised? The burden of proving a job analysis was sufficiently comprehensive is difficult to meet. There is at present no definition of what constitutes a "comprehensive job analysis." The JV process highlights three key elements: (1) persons carrying out the analysis should be job experts; (2) a team approach using independent judgements is most desirable and (3) both frequency and importance of elements should be identified. Varying quality of the job analysis will have a varying effect on V.

If there has been no job analysis of substance in the formulating of the performance appraisal a decision point is reached. If the employer will incur no monetary liability in the form of back pay for past use of the appraisals, then they should be discontinued and a new appraisal system formulated based upon proper job analysis. If potential back pay awards are large, then it may be necessary to move ahead and try to prove they were valid. The probability of establishing the ratings as valid given moderate to severe adverse impact would be low (V = -7 or -10 at this point).

An interesting legal question on job analysis remains largely unanswered. Suppose a defendent in a Title VII litigation involving a performance appraisal system presented a job analysis done post hoc. Further, suppose that the job analysis so done supported the appraisal system. All the guidelines currently endorsed say the job analysis must precede test development. Clearly this requirement is based on the belief that a thorough job analysis is more likely to result in a valid test. But the guidelines say nothing about the evidensary value of a post hoc job analysis. A job analysis is not a sufficient condition for a test to be valid. The question raised here is "Is it necessary?" given the the above suppositions?

The next step in JV involves a consideration of the performance appraisal process. The model indicates that standardized situations are preferable to on the job ratings by supervisors. Further, if the performance is reliably measured by actual output then with almost any type of job analysis the system would be considered valid. An example of such a test for machinists was reported by Schmidt et al (1975).

The next decision point involves an analysis of the structure of the rating form used by the interviewer or supervisor. Ratings of behaviors are considered a plus and rating of traits considered a minus due to the varying degree of subjectivity and level of inference involved in the rater's judgement. If the behaviors are evaluated using behaviorally anchored scales (Campbell, et al, 1973) then V is further enhanced.

The next section indicates an aspect of JV that is unique to performance appraisal systems. Since appraisals involve evaluation of one person by another, the rater must come under scrutiny. The first question asks how many raters are involved. Courts view supervisor judgements as inherently subjective and regard them with great suspicion if they support decisions with adverse impact. In human judgements, there is, to some extent, objectivity in numbers. The extent of training raters receive can adversely affect V if none is given or enhance V if the training is thorough and interrater agreement established at some acceptable level. Finally, the rater's qualifications will influence the outcome of JV. Both experience on the job being evaluated and experience as a rater can enhance the rater's credibility.

Sometimes raters and work sample simulations are combined as in the study of telephone operators by Gail et al (1975) or the study of firefighters we conducted in Baltimore (Livingston, et al 1973).

The next step asks what type of validity is claimed for the appraisal systems. Here the terms used in the guidelines come into play. The main thrust of the flow-chart is that a number of key rational questions relevant to the final decision on validity have been raised before the empirical data are consulted. It is in the context of these rational questions that the weight to be given the empirical evidence is determined.

If a criterion related validity study is available the questions raised in this section involve technical adequacy of the methodology.

At this point familiar psychometric issues are involved. Key points of concern to the courts have been the nature of the sample. As noted earlier separate validity studies for minority groups are definitely preferred. Absence of a separate study for minorities is not a liability if the employer can convince the court such a study was not technically feasible. In the presence of moderate of severe adverse impact, it will be difficult to make such a showing.

The main statistical outputs of a criterion-related validity study are means, standard deviation, correlation coefficients and regression equations. The details of what constitutes a nondiscriminatory test in terms of these various statistics are extremely technical as illustrated by the special issue of the Journal of Educational Measurement last spring (1976). To compound matters, there is considerable differences of opinion among psychologists on the technical definition of a "fair test." It is little wonder that after a court has heard two psychologists express three opinions, they view such "expert opinion" with a jaundiced eye. The following quote from the judge in U S A v. City of Chicago reflects this cynicism:

> The defendants have chosen to lead the court 'deep into the jargon of psychological testing.' The result has been a virtual morass of competing theories advanced by professional testors and tests in which the debate has centered on predictive, concurrent, criterion and construct validation and the court has been left with the unwelcomed task of testing the testers. It is not amiss to observe that plaintiffs have not shunned the debate. (8 EPD 9785)

Some professionals (for example Sharf, 1975) believe the problem can be eliminated by educating the public and the courts. Others feel psychologists need training in giving testimony as experts. I do not feel either approach gets at the heart of the problem. Perhaps the problem is best summarized in the statement by Pogo "We have met the enemy...and they is Us."

I believe the current state of affairs in the litigation of testing is the result of eager endorsement of psychometric ideals by many industrial psychologists who welcomed the judicial review of Title VIII litigation as a means to improve testing practices in business and government. They tried to use the legal process to force changes not easily brought about professionally. But after 8 to 10 years of seeing these psychometric ideals applied via an advocacy process to inferior testing programs many are fearing a monster has been created. A number of the professionals involved early in the process are finding it difficult to back away from the characterization current legal advocates have given to their initial professional opinions. Differential validity is but one example of a concept which enjoyed wide professional endorsement initially but has fallen into disrepute. The theories supporting differential validity still play a large role in litigation. Yet, considering the scientific evidence supporting the concept I believe it fair to say that if "differential validity" were a test it would be enjoined from further use. In short, we have promised the courts, with our high sounding jargon and our sophisticated mathematics, more certainty than we can deliver and are paying the toll for overselling the product. But this is off the subject of how courts have actually viewed statistics which is the primary concern of the present discussion.

The main concern of the courts to date has been with both the statistical and practical significance of the validity coefficient--the correlation between the "test" being validated and the criterion measure of job performance. Statistical significance of a correlation coefficient is a

precisely defined entity. It is determined by comparing the obtained coefficient with a critical value obtained from a table in the back of any statistics book, any person capable of using the logical operators "greater than" or "less than" can determine if a correlation coefficient is statistically significant at some specified level of probability. (The EEOC guidelines specify a less than 1 in 20 probability of chance occurrence as statistically significant). Most modern statistical computer programs provide the significance level to four decimal places on the output. It is only in the significant a level of the correlation coefficient that the validity evidence is "empirical" rather than "rational." It is at this point that the illusion of mathematical objectivity is so misleading. The previous steps in the flow chart emphasize the importance of nonquantitative or "rational" judgements in the total process. An important article entitled "Trial by Mathematics: Precision and Ritual in the Legal Process" (Tribe, 1971) is helpful in gaining a broader perspective on how the apparent elegance and objectivity of mathematical evidence can distort the judicial process. In the case of title VII litigation, Pearson product moment correlations and regression equations have become the mathematical "tail" wagging the judicial "dog."

The overarching importance of rational judgement even at this most empirical step of the process I have called judicial validation is reflected in the concept of "practical significane" articulated by the EEOC guidelines in section 1607.5 (c) (2). Briefly, practical significane is a function

to selection ratios (proportion of applicants actually hired, promoted, or laid off), success ratios (proportion of applicants successful on job without using the test), and business necessity (economic or human risk factors). Of these three factors, selection ratios and business necessity have been the ones involved in the judicial process. I have formulated a three dimensional table intended to characterize the relative influence of various combinations of these factors on V. A situation involving high business necessity, low selection ratios, and a criterion-predictor correlation greater than .30 is the strongest empirical evidence for the validity of the test.

However, assessing practical utility is not the end. The statistical evidence is weighed in terms of the criterion or criteria used in the study. It is at this point that we see the illusion of empirical objectivity most clearly. Because the criteria in the validity study are themselves subject to the scrutiny of judicial validity. At this point the flow-chart loops back to step 2 where the job-relatedness of the criterion is examined. As long as criterion-related data are presented, the looping process will go on; the criterion measure is always subject to rational scruitiny. A point must be reached in the decision process where a criterion is evaluated on its own merit rationally if a decision is to be made. This involves a judgement of its content validity. Thus, while the EEOC guidelines and courts explicitly endorse empirical validity, both logically and realistically, rational analysis is the overarching, more pervasive characteristic of judicial validity. What is portrayed in the flow-chart as judicial validity is similar to the model of "procedural job relatedness"

presented by Kohls (1975). Kohls technique emphasizes consideration of adverse impact, job analysis, use of job experts, and job performance of selected individuals. Judicial validity as presented here is broader in scope.

Since the court must at some point make a judgment concerning content validity we need to look at the key element of content validity. The central element of all available content validity definitions is the requirement that the "test" sample essential behaviors. The only type of measure that satisfies this requirement unambiguously is a job performance simulations. Such a measure has what Guion has labeled "operational validity" (Guion, 1975). The behaviors operationalized in simulation should be the critical behaviors required for job success.

Job simulations are preferrable to actual on the job behaviors in several respects. The standardization of task demands allows for control of exogenous influnces on ratee performance. Quantification of performance can be objectified even if raters are involved, through training and use of multiple raters. On-the-job ratings by a single supervisor, even when behaviorally anchored scales are employed, is suspect in court due to the "inherent" subjectivity of a personal judgement. Recent research by Gael, Grant, & Ritchie (1975a, 1975b) present two job performance measures which possess a high level of operational validity. Even though these measures have marked adverse impact (minorities performed significantly lower than nonminorities on all dimensions of both job simulations) it is my judgement that they would be found job-related within the judicial validation process. Assume that the authors had only used behaviorly anchored scales to obtain supervisor ratings of on-the-job performance. Extrapolating from the

simulation results reported by the authors, the supervisor ratings would also evidence adverse impact. Given the extent of adverse impact and court skepticism of "subjective" supervisor ratings, the status of the ratings within judicial validity would be questionable. The relationship between the behaviorally anchored supervisor ratings of on-the-job performance and job-simulation performance would need to be demonstrated.

Social Utility

Civil rights and employment testing involves social, emotional and philosophical as well as the job related issues. To those individuals who are especially sensitive to these other issues my presentation may appear crassly pragmatic. If this is so, it is because I have limited my presentation to issues explicitly addressed in the federal guidelines or the case.

There are social questions raised by Title VII which affects our individual view points. Critics of employment testing are committed to increasing minority participation in all levels of the work force. Many endorse preferential treatment of minorities to accomplish this goal stating "race conscious evils require race conscious remedies." The recent Supreme Court decision to dodge this issue in the DeFunis reverse discrimination case involving a University of Washington law student reflects our societal reluctance to grapple with the issue of social utilities explicitly at any policy level.

I would argue that the judicial validity model reflects the courts position on at least three factors influencing social utility. These factors are adverse impact, selection ratios, business necessity. The

adverse impact factor is loaded in favor of minorities. The courts have made validity requirements more stringent as a function of adverse impact. The other two factors attempt to consider utility from the employers point of view. What is not reflected in the model is the social utility of the individual judges. Anyone who has been involved in several court cases is struck by the wide variation in behavior of judges. My own inference is that the judges' behavior in handling Title VII cases is significantly influenced by their personal social utilities. The appelate system of courts is supposed to offset a judge's personal biases. But the first judge at the district level has a great deal of leeway in trial procedure while still avoiding reversable error. Judicial utility is an unknown factor until the case is actually assigned to a judge, but it will have an influence.

My personal opinion is that the social utilities which have evolved from the judicial process as noted earlier are reasonable. Furthermore, the judicial system is the only social policy force in our current system which could systematically explore such a controversial area unfettered by the power of special interest which distort our cultural values and immobilize our legislative system.

REFERENCES

Albemarle Paper Co. v Moody 422 U.S. (1975)

American Psychological Association. Standards to educational and psychological tests (Rev. Ed.). Washington D. C.: Author, 1974.

"Application of the EEOC Guidelines to Employment Test Validation: A Uniform Standard for Both Public and Private Employers." George Washington Law Review, 1973, 41, 505-518.

Ash, P. "Implications of the Civil Rights Act of 1964 for psychological assessment in industry." American Psychologist, 1966, 21, 797-803.

Campbell, J. P. & Dunnette, M. D., Arvey, R. D. Heller "The development and evaluation of behaviorally based rating scales." Journal of Applied Psychology, 57, 1973, 15-22.

Gael, S., Grant, D. L., & Ritchie, R. J. "Employment test validation for minority and nonminority telephone operators". Journal of Applied Psychology, 1975, 60, No. 4, 411-419.

Gael, S., Grant, D. T., & Ritchie, R. "Employment Test Validation for Minority and Nonminority Clerks with Work Sample Criteria". Journal of Applied Psychology, 1975, 60, 420-426.

Guion, R. "Content validity, the source of my discontent." A paper presented at the Amer. Psych. Assoc. Convention, Chicago, Sept. 1975.

Griggs v. Duke Power Co., 401 US (1971)

Harper v. Mayor and City Council of Baltimore. 359 F Supp. (1973)

Holley, W. H. & Fields, H. S. Performance appraisal and the law. Labor Law Journal, 1975, July, 423-430.

Kohls, J. W. Equal employment opportunity guidelines and validation: A new approach to job-relatedness. Paper presented at the meeting of the American Psychological Association, Chicago, Sept. 1975.

Lazer, Robert I. The discrimination danger in performance appraisal. The Conference Board Record, 1976, XIII, No. 3, 60-64.

Livingston, S. A., Edwards, K. J., & Wright, C. S. Valicity of an examination for the selection of fire fighter trainees. A report to the Civil Service Commission of Baltimore, Maryland October, 1973.

Miner, J. B. Psychological testing and fair employment practices; A testing program that does not discriminate. Personnel Psychology, 1974, 27, 49-62.

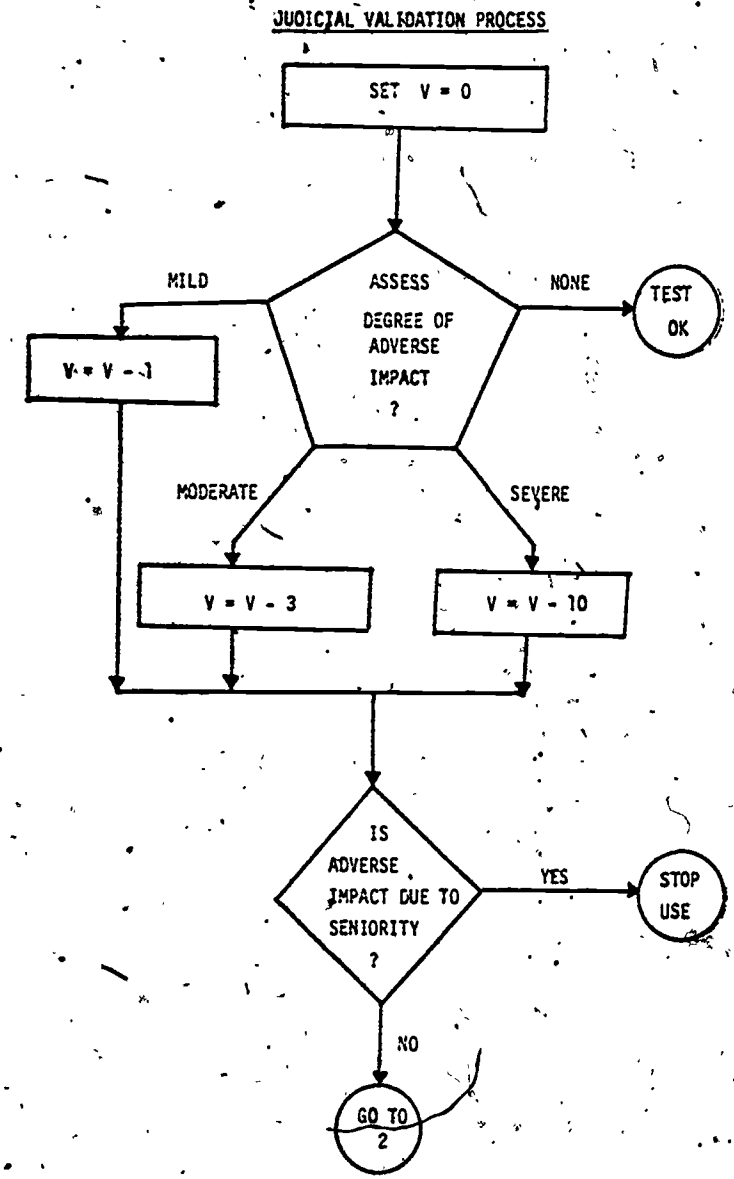
Peterson, N. S. & Novick, M. R. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

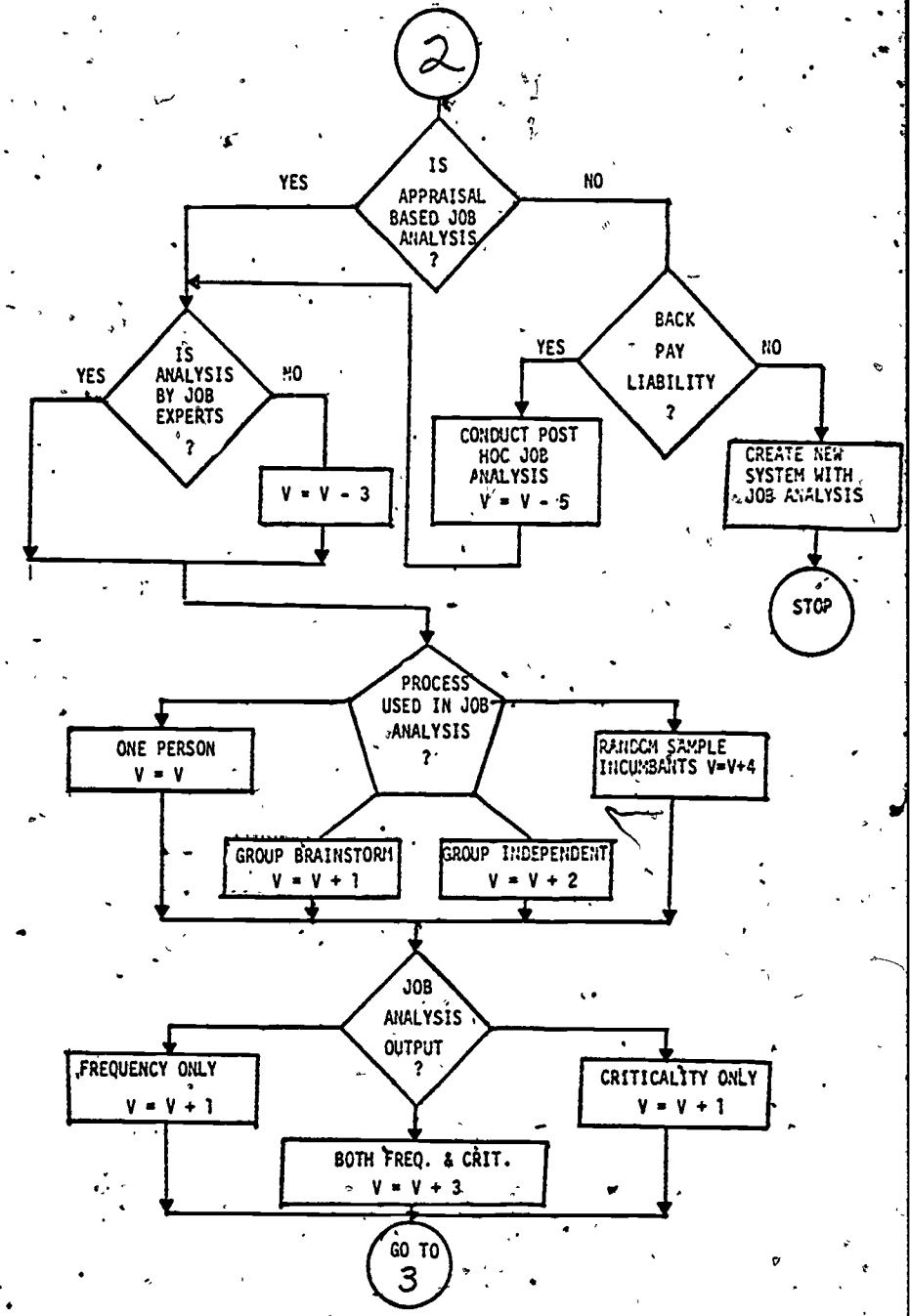
Scharf, James C. Influence of Lawyers, Legal Language and Legal Thinking. A paper presented at the Annual Convention of the American Psychological Association, Chicago, Sept. 1, 1975.

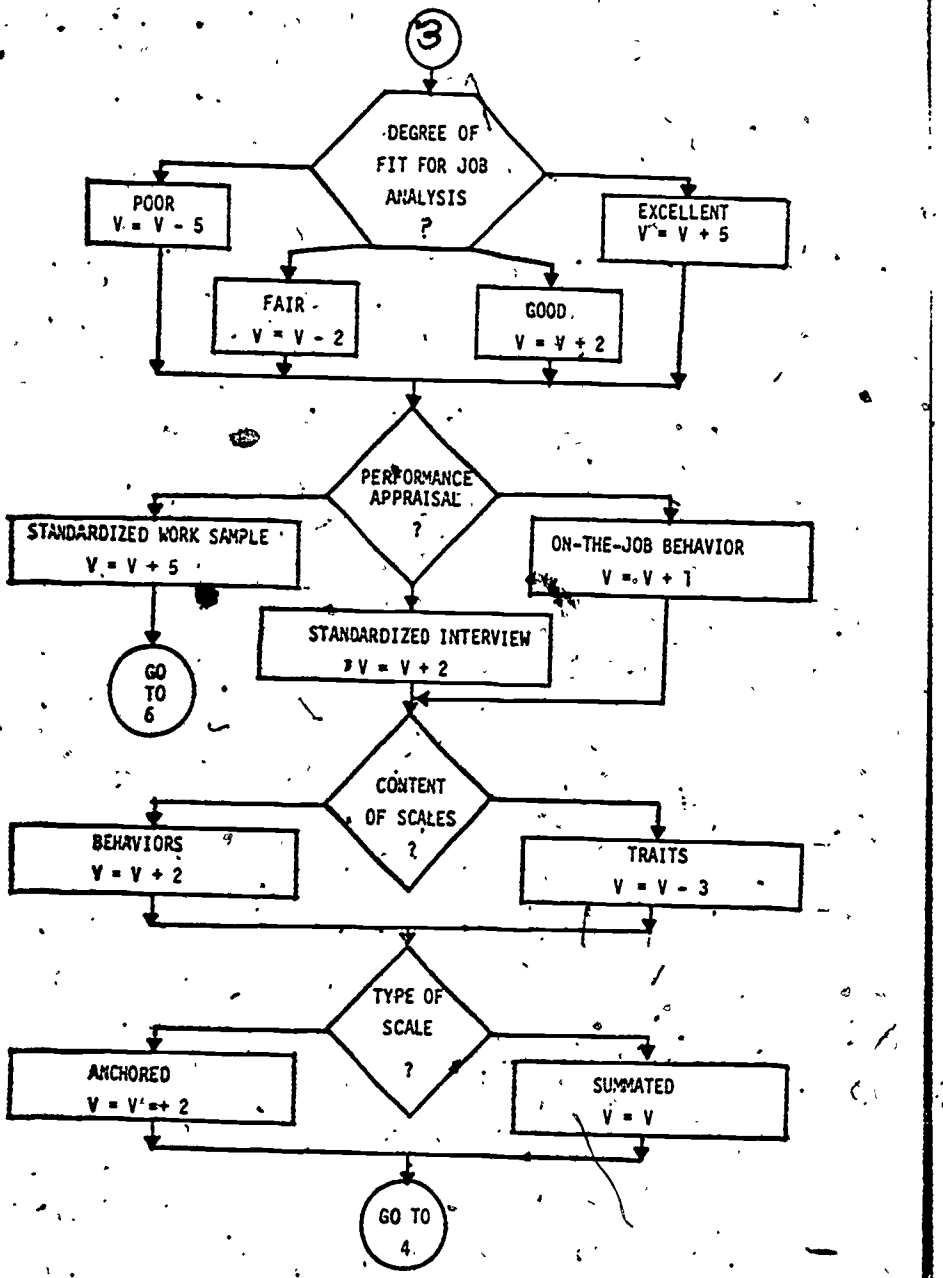
Schmidt, F. L., Greenthal, A. L., Berner, J. G., Hunter, J. E., & Seaton, F. W. Job sample vs paper-and-pencil trades and technical tests: adverse impact and examinee attitudes. A paper presented at the Annual Convention of the American Psychological Association, September, 1975, Chicago, Ill.

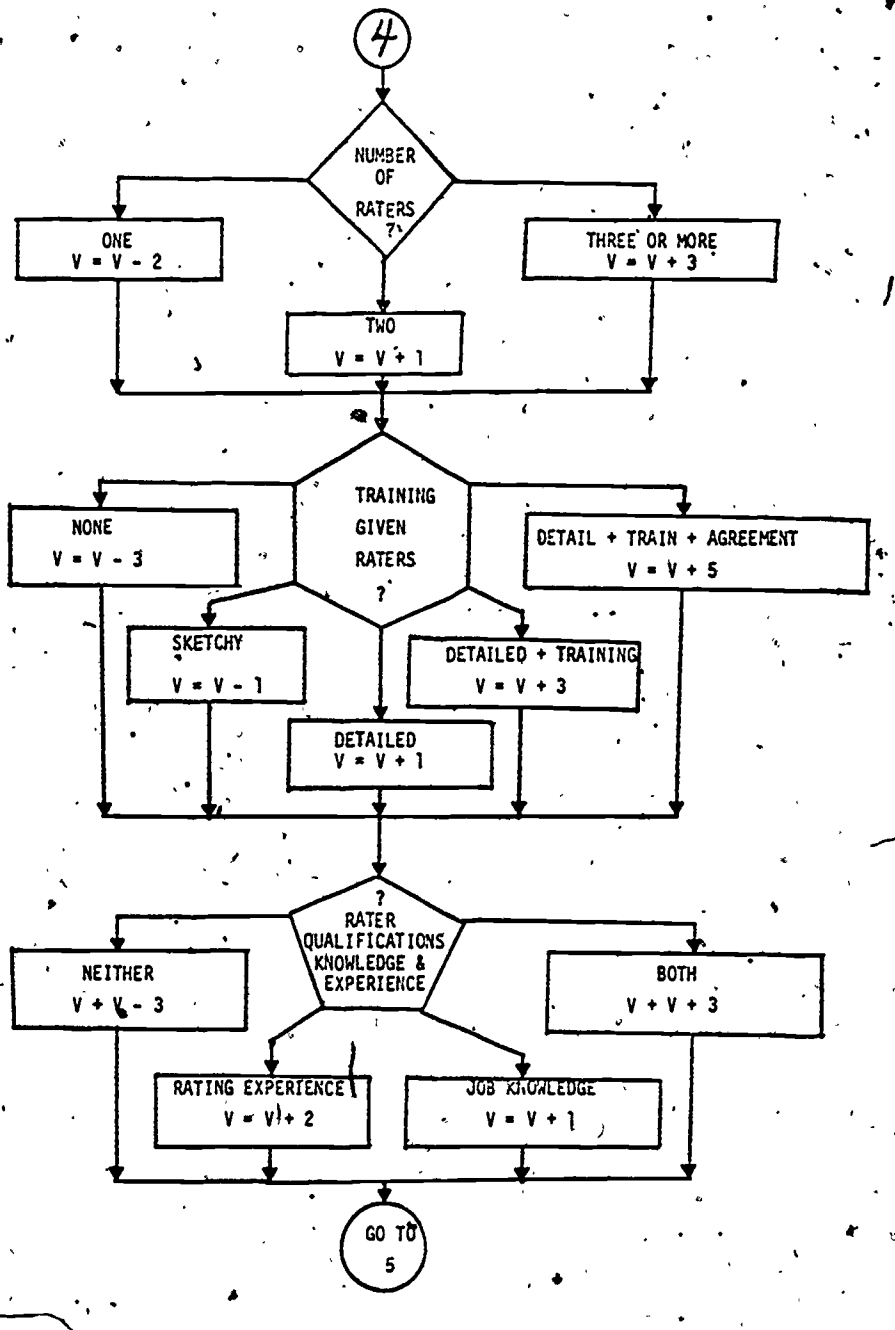
Tribe, L. H. Trial by mathematics: precision and ritual in the legal process. Harvard Law Review, 1971, 81, 1329-1393.

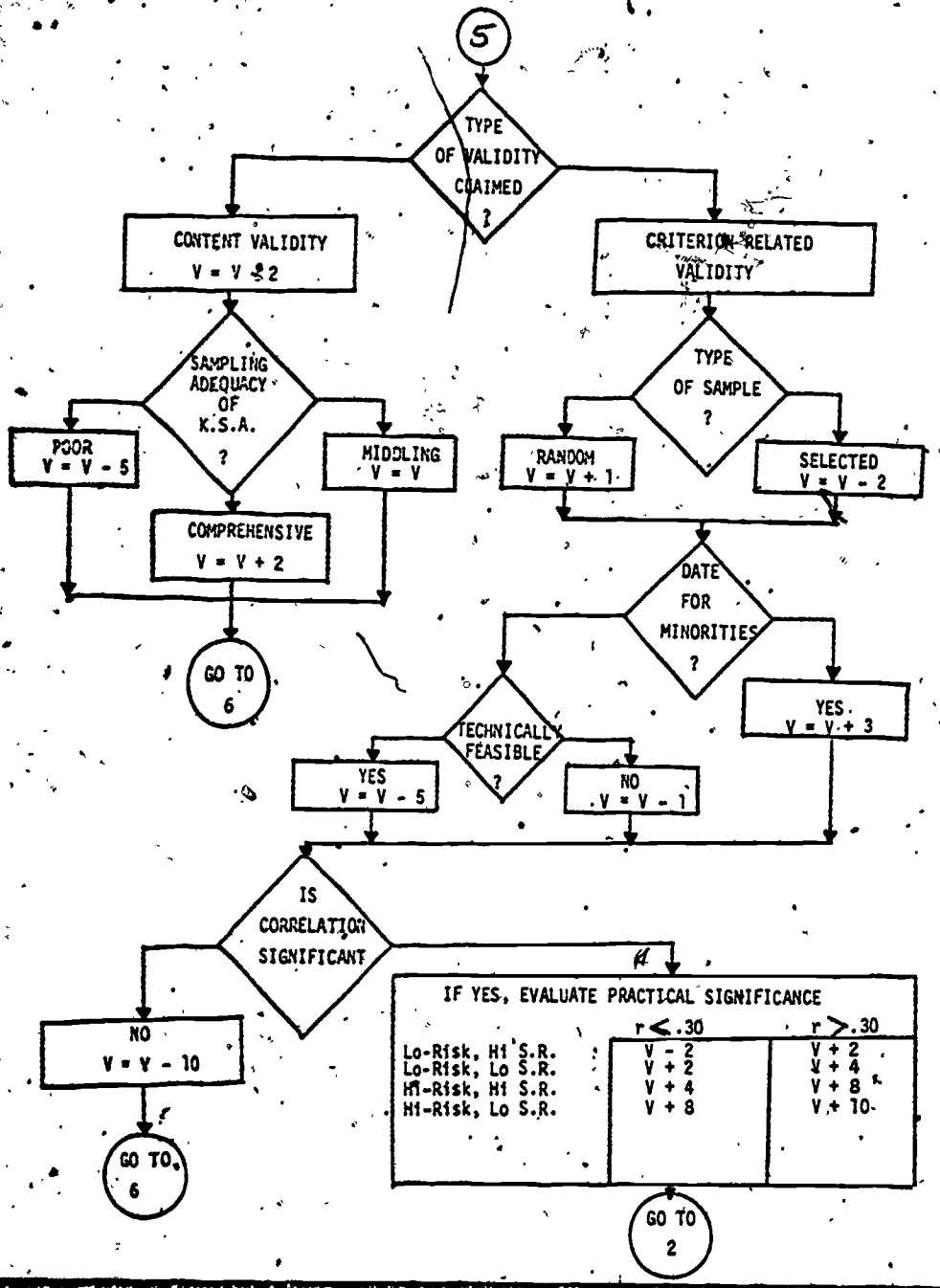
U.S.A. v. City of Chicago, 8 EPD (1976)

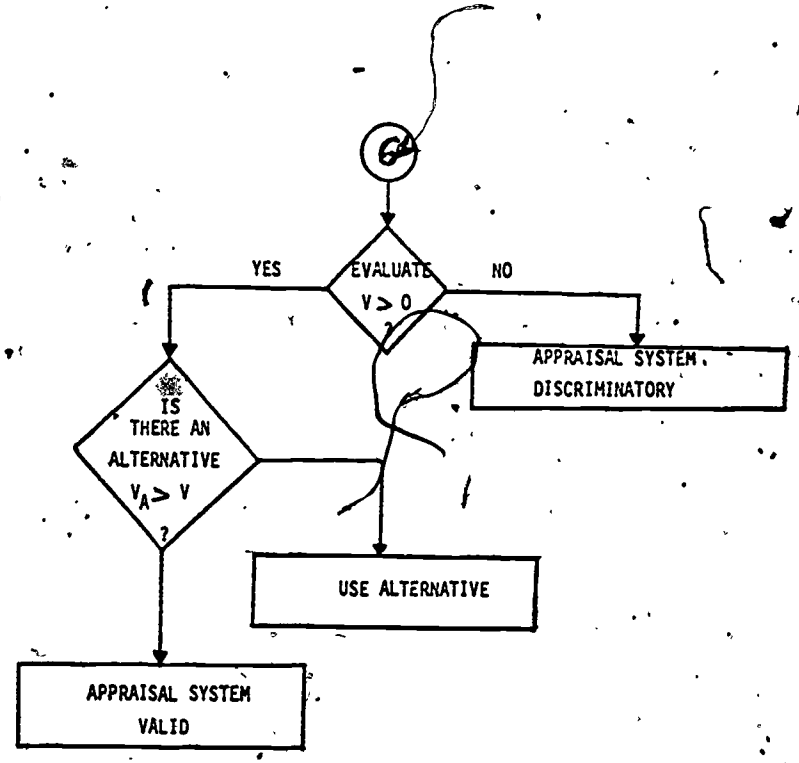

JUDICIAL VALIDATION PROCESS
SET V = 0
ASSESS DEGREE OF ADVERSE IMPACT ?
MILD
NONE
TEST OK
V = V - 1
MODERATE
SEVERE
V = V - 3
V = V - 10
IS ADVERSE IMPACT DUE TO SENIORITY ?
YES
STOP USE
NO
GO TO 2


2
IS APPRAISAL BASED JOB ANALYSIS ?
YES
NO
IS ANALYSIS BY JOB EXPERTS ?
YES
NO
V = V - 3
BACK PAY LIABILITY ?
YES
NO
CONDUCT POST HOC JOB ANALYSIS V = V - 5
CREATE NEW SYSTEM WITH JOB ANALYSIS
STOP
PROCESS USED IN JOB ANALYSIS ?
ONE PERSON V = V
RANDOM SAMPLE INCUMBANTS V=V+4
GROUP BRAINSTORM V = V + 1
GROUP INDEPENDENT V = V + 2
JOB ANALYSIS OUTPUT ?
FREQUENCY ONLY V = V + 1
CRITICALITY ONLY V = V + 1
BOTH FREQ. & CRIT. V = V + 3
GO TO 3


3
DEGREE OF FIT FOR JOB ANALYSIS ?
POOR V = V - 5
FAIR V = V - 2
GOOD V = V + 2
EXCELLENT V = V + 5
PERFORMANCE APPRAISAL ?
STANDARDIZED WORK SAMPLE V = V + 5
ON-THE-JOB BEHAVIOR V = V + 1
STANDARDIZED INTERVIEW V = V + 2
GO TO 6
CONTENT OF SCALES ?
BEHAVIORS V = V + 2
TRAITS V = V - 3
TYPE OF SCALE ?
ANCHORED V = V + 2
SUMMATED V = V
GO TO 4


4
NUMBER OF RATERS ?
ONE V = V - 2
TWO V = V + 1
THREE OR MORE V = V + 3
TRAINING GIVEN RATERS ?
NONE V = V - 3
SKETCHY V = V - 1
DETAILED V = V + 1
DETAILED + TRAINING V = V + 3
DETAIL + TRAIN + AGREEMENT V = V + 5
RATER QUALIFICATIONS KNOWLEDGE & EXPERIENCE ?
NEITHER V = V - 3
RATING EXPERIENCE V = V + 2
JOB KNOWLEDGE V = V + 1
BOTH V = V + 3
GO TO 5


5
TYPE OF VALIDITY CLAIMED ?
CONTENT VALIDITY
V = V ± 2
CRITERION RELATED VALIDITY
SAMPLING ADEQUACY OF K.S.A. ?
POOR
V = V - 5
MIDDLING
V = V
COMPREHENSIVE
V = V + 2
TYPE OF SAMPLE ?
RANDOM
V = V + 1
SELECTED
V = V - 2
GO TO 6
DATE FOR MINORITIES ?
YES
V = V + 3
TECHNICALLY FEASIBLE ?
YES
V = V - 5
NO
V = V - 1
IS CORRELATION SIGNIFICANT
NO
V = V - 10
GO TO 6
IF YES, EVALUATE PRACTICAL SIGNIFICANCE
r ≤ .30
r > .30
Lo-Risk, Hi S.R.
V - 2
V + 2
Lo-Risk, Lo S.R.
V + 2
V + 4
Hi-Risk, Hi S.R.
V + 4
V + 8
Hi-Risk, Lo S.R.
V + 8
V + 10
GO TO 2


6
EVALUATE V > 0 ?
YES
NO
APPRAISAL SYSTEM DISCRIMINATORY
IS THERE AN ALTERNATIVE $V_A > V$?
USE ALTERNATIVE
APPRAISAL SYSTEM VALID